Biostat 200C Homework 1

Due Apr 16 @ 11:59PM

Zian ZHUANG

Q1. Binomial Distribution

Let Y_i be the number of successes in n_i trials with

$$Y_i \sim Bin(n_i, \pi_i),$$

where the probabilities π_i have a Beta distribution

$$\pi_i \sim Beta(\alpha, \beta).$$

The probability density function for the Beta distribution is $f(x; \alpha, \beta) = x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha, \beta)$ for $x \in [0,1], \alpha > 0, \beta > 0$, and the beta function $B(\alpha, \beta)$ defining the normalizing constant required to ensure that $\int_0^1 f(x; \alpha, \beta) = 1$. Let $\theta = \alpha/(\alpha+\beta)$, show that

a. $E(\pi_i) = \theta$

$$\begin{aligned}
E(\pi_i) &= \int \pi_i * f(\pi_i) d\pi_i \\
&= \int \pi_i * \pi_i^{\alpha-1}(1-\pi_i)^{\beta-1}/B(\alpha,\beta) d\pi_i \\
&= B(\alpha,\beta)^{-1} \int \pi_i^{(\alpha+1)-1}(1-\pi_i)^{\beta-1} d\pi_i \\
&= B(\alpha+1,\beta) * B(\alpha,\beta)^{-1} \int B(\alpha+1,\beta)^{-1} * \pi_i^{(\alpha+1)-1}(1-\pi_i)^{\beta-1} d\pi_i \\
&= B(\alpha+1,\beta) * B(\alpha,\beta)^{-1} * 1 \\
&= \Gamma(\alpha+1)\Gamma(\beta)/\Gamma(\alpha+1+\beta)/(\Gamma(\alpha)\Gamma(\beta)) * \Gamma(\alpha+\beta) \\
&= \alpha/(\alpha+\beta) \\
&= \theta
\end{aligned}$$

b. $Var(\pi_i) = \theta(1-\theta)/(\alpha+\beta+1) = \phi\theta(1-\theta)$ Firstly we can calculated $E(\pi_i^2)$

$$\begin{aligned}
E(\pi_i^2) &= \int \pi_i^2 * f(\pi_i) d\pi_i \\
&= \int \pi_i^2 * \pi_i^{\alpha-1}(1-\pi_i)^{\beta-1}/B(\alpha,\beta) d\pi_i \\
&= B(\alpha,\beta)^{-1} \int \pi_i^{(\alpha+2)-1}(1-\pi_i)^{\beta-1} d\pi_i \\
&= B(\alpha+2,\beta) * B(\alpha,\beta)^{-1} \int B(\alpha+1,\beta)^{-1} * \pi_i^{(\alpha+2)-1}(1-\pi_i)^{\beta-1} d\pi_i \\
&= B(\alpha+2,\beta) * B(\alpha,\beta)^{-1} * 1 \\
&= \Gamma(\alpha+2)\Gamma(\beta)/\Gamma(\alpha+2+\beta)/(\Gamma(\alpha)\Gamma(\beta)) * \Gamma(\alpha+\beta) \\
&= \alpha * (\alpha+1)/(\alpha+1+\beta) * (\alpha+\beta) \\
&= \theta(\alpha+1)/(\alpha+1+\beta)
\end{aligned}$$

Then we can obtain $Var(\pi_i)$

$$\begin{aligned}
Var(\pi_i) &= E(\pi_i^2) - E(\pi_i)^2 \\
&= ((\alpha+1)\alpha(\alpha+\beta) - \alpha^2(\alpha+\beta+1))/(\alpha+\beta+1)(\alpha+\beta)^2 \\
&= (\alpha\beta)/(\alpha+\beta)^2/(\alpha+1+\beta) \\
&= \theta(1-\theta)/(\alpha+\beta+1) = \phi\theta(1-\theta)
\end{aligned}$$

c. $E(Y_i) = n_i\theta$

$$\begin{aligned}
E(Y_i) &= E_{\pi_i}(E_{Y_i}(Y_i|\pi_i)) \\
&= E_{\pi_i}(n_i * \pi_i) \\
&= n_i * E(\pi_i) \\
&= n_i * \theta
\end{aligned}$$

d. $Var(Y_i) = n_i\theta(1-\theta)[1+(n_i-1)\phi]$ so that $Var(Y_i)$ is larger than the Binomial variance (unless $n_i = 1$ or $\phi = 0$).

$$\begin{aligned}
Var(Y_i) &= E_{\pi_i}(Var(Y_i|\pi_i)) + Var_{\pi_i}(E(Y_i|\pi_i)) \\
&= E_{\pi_i}(n_i * \pi_i * (1-\pi_i)) + Var_{\pi_i}(\pi_i * n_i) \\
&= n_i * (E(\pi_i) - E(\pi_i^2)) + n_i^2 * \phi\theta(1-\theta) \\
&= n_i * (\theta - \theta(\alpha+1)/(\alpha+1+\beta)) + n_i^2 * \phi\theta(1-\theta) \\
&= n_i * (\theta(1-(\alpha+1)/(\alpha+1+\beta))) + n_i^2 * \phi\theta(1-\theta) \\
&= n_i * (\theta * \beta/(\alpha+1+\beta)) + n_i^2 * \phi\theta(1-\theta) \\
&= n_i * (\theta * (1-\theta)(1-\phi)) + n_i^2 * \phi\theta(1-\theta) \\
&= n_i\theta(1-\theta)[1+(n_i-1)\phi]
\end{aligned}$$

Q2. (ELMR Chapter 3 Exercise 1)

A case-control study of esophageal cancer in Ileet-Vilaine, France.

```
data(esoph)
```

a. Plot the proportion of cases against each predictor using the size of the point to indicate the number of subject as seen in Figure 2.7. Comment on the realtionships seen in the plots.

Solution:

b. Fit a binomial GLM with interactions between all three predictors. Use AIC as a criterion to select a model using the `step` function. Which model is selected?

Solution:

c. All three factors are ordered and so special contrasts have been used appropriate for ordered factors involving linear, quadratic and cubic terms. Further simplification of the model may be possible by eliminating some of these terms. Use the `unclass` function to convert the factors to a numerical representation and check whether the model may be simplified.

Solution:

d. Use the summary output of the factor model to suggest a model that is slightly more complex than the linear model proposed in the previous question.

Solution:

e. Does your final model fit the data? Is the test you make accurate for this data?

Solution:

f. Check for outliers in your final model.

Solution:

g. What is the predicted effect of moving one category higher in alcohol consumption?

Solution:

h. Compute a 95% confidence interval for this predicted effect.

Solution: